On the Theory of Correlation for any Number of Variables, treated by a New System of Notation.

By G. Udny Yule, Newmarch Lecturer on Statistics, University College, London.

(Communicated by Professor O. Henrici, F.R.S. Received January 25,—Read February 28, 1907.)

1. The systems of notation hitherto used by writers on the theory of correlation are somewhat unsatisfactory when many variables are involved. In the present paper a new notation is proposed which is simple, definite, and quite general, thus very greatly facilitating the treatment of the subject. The majority of the results given in the sequel were, in fact, first suggested by the notation itself.

2. Let $x_1x_2...x_n$ denote deviations in the values of the n variables from their respective arithmetic means. Then the regression equation may be written :—

$$x_1 = b_{12\cdot34...n}x_2 + b_{13\cdot24...n}x_3 + \ldots + b_{1n\cdot23...n-1}x_n. \quad (1)$$

In this notation the suffix of each regression coefficient completely defines it. The first subscript gives the dependent variable, the second the variable of which the given regression is the coefficient, and the subscripts after the period show the remaining independent variables which enter into the equation. It is convenient to distinguish the subscripts before and after the period as "primary" and "secondary" subscripts respectively. The order in which the secondary subscripts are arranged is indifferent, but the order of the two primary subscripts is material; *e.g.*, $b_{12\cdot3...n}$ and $b_{21\cdot3...n}$ denote two quite distinct coefficients. A coefficient with p secondary subscripts may be termed a regression of the pth order, the total regressions b_{12}, b_{13}, b_{23}, etc., being thus regarded as of order zero.

3. The correlation-coefficients may be distinguished by subscripts in precisely the same manner. Thus the correlation $r_{12\cdot34...n}$ is defined by the relation

$$r_{12\cdot34...n} = (b_{12\cdot34...n} \cdot b_{21\cdot34...n})^{\frac{1}{2}}. \quad (2)$$

In the case of the correlations, the order of both primary and secondary subscripts is indifferent. A correlation with p secondary subscripts may be termed a correlation of order p, the total correlations r_{12}, r_{13}, r_{23}, etc., being regarded as of order zero.

4. If the regressions in equation (1) be determined as usual by the method

of least squares, the difference between x_1 and the expression on the right, for any observed set of values of $x_1x_2...x_n$, may be denoted by $x_{1\cdot23...n}$: that is

$$x_{1\cdot23...n} = x_1 - b_{12\cdot34...n}x_2 - b_{13\cdot24...n}x_3 - \ldots - b_{1n\cdot23...n-1}x_n. \tag{3}$$

Such a residual, or deviation, denoted by a symbol with p secondary subscripts may be termed a deviation of the pth order, $x_1x_2...x_n$ being regarded as deviations of order zero.

5. Finally, the standard deviation $\sigma_{1\cdot23...n}$ is defined as given by the relation

$$\mathrm{N}\,.\,\sigma^2_{1\cdot23...n} = \Sigma\,(x^2_{1\cdot23...n}), \tag{4}$$

N being the number of observations. If the standard deviation be denoted by a symbol with p secondary subscripts, it is of the pth order, the total standard deviations being regarded as of order zero.

6. In terms of this notation, the normal equations from which the regressions are determined may be very briefly written, in the form

$$\begin{aligned}\Sigma\,(x_2\,.\,x_{1\cdot23...n}) = \Sigma\,(x_3\,.\,x_{1\cdot23...n}) &= \ldots \\ = \Sigma\,(x_n\,.\,x_{1\cdot23...n}) &= 0.\end{aligned} \tag{5}$$

That is to say, we have the general theorem: "The product-sum of any deviation of order zero with any deviation of higher order is zero, provided the subscript of the former occur amongst the secondary subscripts of the latter."

7. It follows that the product-sum of any two deviations of the same order, with the same secondary suffixes, is unaltered by omitting any or all of the secondary subscripts of either and, conversely, the product-sum of any deviation of order p with a deviation of order $p+q$, the p subscripts being the same in each case, is unaltered by adding to the secondary subscripts of the former any or all of the q additional subscripts of the latter, for we have by § 6:—

$$\begin{aligned}\Sigma\,(x_{1\cdot34...n}x_{2\cdot34...n}) &= \Sigma\,(x_{1\cdot34...n})\,(x_2 - b_{23\cdot4...n}x_3 - \ldots - b_{2n\cdot3...n-1}x_n) \\ &= \Sigma\,(x_{1\cdot34...n}x_2).\end{aligned}$$

Similarly

$$\Sigma\,(x_{1\cdot34...n}x_{2\cdot34...n-1}) = \Sigma\,(x_{1\cdot34...n}x_2),$$

and so on. Therefore, quite generally,

$$\begin{aligned}\Sigma\,(x_{1\cdot34...n}x_{2\cdot34...n}) = \Sigma\,(x_{1\cdot34...n}x_{2\cdot34...n-1}) &= \ldots \\ \ldots = \Sigma\,(x_{1\cdot34...n}x_2).\end{aligned} \tag{6}$$

8. It follows from § 7 as a corollary from § 6 that the product-sum of any two deviations is zero if all the subscripts of the one are contained among the secondary subscripts of the other.

These theorems (§§ 6—8) give the key to simple deductions of many results in the theory of multiple correlation.

9. We have from the last section and § 7,

$$\begin{aligned} 0 &= \Sigma(x_{2\cdot34\ldots n}x_{1\cdot234\ldots n}) \\ &= \Sigma(x_{2\cdot34\ldots n})(x_1 - b_{12\cdot34\ldots n}x_2 - \text{terms in } x_3 \text{ to } x_n) \\ &= \Sigma(x_1x_{2\cdot34\ldots n}) - b_{12\cdot34\ldots n}\Sigma(x_2x_{2\cdot34\ldots n}) \\ &= \Sigma(x_{1\cdot34\ldots n}x_{2\cdot34\ldots n}) - b_{12\cdot34\ldots n}\Sigma(x^2{}_{2\cdot34\ldots n}). \end{aligned}$$

That is

$$b_{12\cdot34\ldots n} = \frac{\Sigma(x_{1\cdot34\ldots n}x_{2\cdot34\ldots n})}{\Sigma(x^2{}_{2\cdot34\ldots n})}. \tag{7}$$

But this is the value that would have been obtained by taking a regression equation of the form

$$x_{1\cdot34\ldots n} = b_{12\cdot34\ldots n}x_{2\cdot34\ldots n},$$

and determining $b_{12\cdot34\ldots n}$ by the method of least squares. That is to say, $b_{12\cdot34\ldots n}$ may be regarded, quite generally and without any reference to the form of the frequency distribution, as the regression of $x_{1\cdot34\ldots n}$ on $x_{2\cdot34\ldots n}$. It follows at once from the definition (3) that $r_{12\cdot34\ldots n}$ may be regarded as the correlation between $x_{1\cdot34\ldots n}$ and $x_{2\cdot34\ldots n}$, and from (4) that we may write

$$b_{12\cdot34\ldots n} = r_{12\cdot34\ldots n}\frac{\sigma_{1\cdot34\ldots n}}{\sigma_{2\cdot34\ldots n}}. \tag{8}$$

All the relations, in fact, that hold good between deviation-sums, standard deviations, regressions and correlations of order zero, are also valid between deviation-sums, standard deviations, regressions and correlations of any high order.

10. This result is of some importance as regards the interpretation of partial correlations and regressions. In the case of normal correlation there is no difficulty in assigning a meaning to these constants, as the regression is strictly linear, and the partial correlations and regressions are the same for all types of the variables. But in the general case this is not so, and although I showed, in a previous discussion of the question,* that the values assigned to the partial regressions on the assumption of normal correlation are the "least square" values and, consequently, that the partial correlation retains an "average significance," I could not prove that it remains an actual correlation between determinate variables. The above theorem completes the work in this respect. If, with three variables x_1, x_2, and x_3, for example, the two regressions b_{13} and b_{23} be determined in the ordinary way, and then the residuals $x_{1\cdot3} = x_1 - b_{13}x_3$, $x_{2\cdot3} = x_2 - b_{23}x_3$ be calculated for all sets of observations $x'_1x'_2x'_3$, $x''_1x''_2x''_3$, etc., the correlation between $x_{1\cdot3}$ and $x_{2\cdot3}$ is $r_{12\cdot3}$. A similar interpretation holds for any greater number of variables.

* 'Roy. Soc. Proc.,' vol. 60 (1897), p. 477; 'Roy. Stat. Soc. Journ.,' vol. 60 (1897), p. 812.

Such a relation would not, of course, afford a practical method of calculating the partial coefficients, as the arithmetic would be extremely lengthy.

11. Any standard deviation of order p may be expressed in terms of a standard deviation of order $p-1$ and a correlation of order $p-1$. For we have, using the theorems of §§ 6 and 7,

$$\begin{aligned}\Sigma(x^2_{1\cdot 23\ldots n}) &= \Sigma(x_{1\cdot 23\ldots n-1}x_{1\cdot 23\ldots n})\\ &= \Sigma(x_{1\cdot 23\ldots n-1})(x_1 - b_{1n\cdot 23\ldots n-1}x_n - \text{terms in } x_2 \text{ to } x_{n-1})\\ &= \Sigma(x^2_{1\cdot 23\ldots n-1}) - b_{1n\cdot 23\ldots n-1}\Sigma(x_{1\cdot 23\ldots n-1}x_{n\cdot 23\ldots n-1});\end{aligned}$$

or, dividing through by the number of observations,

$$\begin{aligned}\sigma^2_{1\cdot 23\ldots n} &= \sigma^2_{1\cdot 23\ldots n-1}(1 - b_{1n\cdot 23\ldots n-1}b_{n1\cdot 23\ldots n-1})\\ &= \sigma^2_{1\cdot 23\ldots n-1}(1 - r^2_{1n\cdot 23\ldots n-1}).\end{aligned} \tag{9}$$

The form of this relation is the same as that of the familiar relation between a standard deviation of the first order and a standard deviation of order zero, with the secondary subscripts $_{23\ldots n-1}$ added throughout. It is clear from (9) that $r_{1n\cdot 23\ldots n-1}$ cannot be numerically greater than unity. It also follows at once that if we have been estimating x_1 from $x_2, x_3 \ldots x_{n-1}$, x_n will not increase the accuracy of estimate unless $r_{1n\cdot 23\ldots n-1}$ (not r_{1n}) differ from zero.*

12. In equation (9) the subscript n is eliminated from the suffix of $\sigma_{1\cdot 23\ldots n}$, and it is obvious that any other subscript can be eliminated in the same way. Therefore we must have

$$\begin{aligned}\sigma^2_{1\cdot 34\ldots n}(1 - r^2_{12\cdot 34\ldots n}) &= \sigma^2_{1\cdot 24\ldots n}(1 - r^2_{13\cdot 24\ldots n})\\ &= \ldots = \sigma^2_{1\cdot 23\ldots n-1}(1 - r^2_{1n\cdot 23\ldots n-1}).\end{aligned} \tag{10}$$

Further, we have

$$\begin{aligned}\sigma^2_{1\cdot 23\ldots n-1} &= \sigma^2_{1\cdot 23\ldots n-2}(1 - r^2_{1(n-1)\,\cdot\, 23\ldots n-2})\\ \sigma^2_{1\cdot 23\ldots n-2} &= \sigma^2_{1\cdot 23\ldots n-3}(1 - r^2_{1(n-2)\,\cdot\, 23\ldots n-3}),\end{aligned}$$

and so on; so that

$$\sigma^2_{1\cdot 23\ldots n} = \sigma^2_1(1 - r^2_{12})(1 - r^2_{13\cdot 2})(1 - r^2_{14\cdot 23})\ldots(1 - r^2_{1n\cdot 23\ldots n-1}). \tag{11}$$

This is an extremely convenient expression for arithmetical use, as illustrated later. A complete check on the arithmetic is obtained by eliminating the secondary subscripts in a different, say the inverse, order, *i.e.*, by using the result—

$$\sigma^2_{1\cdot 23\ldots n} = \sigma^2_1(1 - r^2_{1n})(1 - r^2_{1(n-1)\,\cdot\, n})(1 - r^2_{1\,(n-2)\,\cdot\, n\,(n-1)})\ldots(1 - r^2_{12\cdot 34\ldots n}). \tag{12}$$

* *Cf.* proofs for cases of 3 and 4 variables previously given (*loc. cit.* in previous note).

13. Any regression of order p may be expressed in terms of regressions of order $p-1$. For we have

$$\begin{aligned}\Sigma(x_{1.34\ldots n}x_{2.34\ldots n}) &= \Sigma(x_{1.34\ldots n-1}x_{2.34\ldots n})\\ &= \Sigma(x_{1.34\ldots n-1})(x_2 - b_{2n.34\ldots n-1}x_n - \text{terms in } x_3 \text{ to } x_{n-1})\\ &= \Sigma(x_{1.34\ldots n-1}x_{2.34\ldots n-1}) - b_{2n.34\ldots n-1}\Sigma(x_{1.34\ldots n-1}x_{n.34\ldots n-1}).\end{aligned}$$

That is, replacing $b_{2n.34\ldots n-1}$ by $b_{n2.34\ldots n-1} \times \sigma^2_{2.34\ldots n-1}/\sigma^2_{n.34\ldots n-1}$,

$$b_{12.34\ldots n}\sigma^2_{2.34\ldots n} = b_{12.34\ldots n-1}\sigma^2_{2.34\ldots n-1} - b_{1n.34\ldots n-1}b_{n2.34\ldots n-1}\sigma^2_{2.34\ldots n-1}.$$

Therefore, by equation (9),

$$b_{12.34\ldots n} = \frac{b_{12.34\ldots n-1} - b_{1n.34\ldots n-1}\,.\,b_{n2.34\ldots n-1}}{1 - b_{1n.34\ldots n-1}b_{n1.34\ldots n-1}}. \qquad (13)$$

But this is simply the expression for $b_{12.n}$ in terms of b_{12}, b_{1n}, b_{n1}, and b_{n2}, with the subscripts $_{34\ldots n-1}$ added throughout. Therefore $b_{12.34\ldots n}$ may be regarded as the partial regression of $x_{1.34\ldots n-1}$ on $x_{2.34\ldots n-1}$, $x_{n.34\ldots n-1}$ being given. As any other secondary subscript might have been eliminated in lieu of n, we can also regard it as the partial regression of $x_{1.45\ldots n}$, on $x_{2.45\ldots n}$, $x_{3.45\ldots n}$ being given, and so on.

14. Equation (13) may be written in terms of the correlations:—

$$b_{12.34\ldots n} = \frac{r_{12.34\ldots n-1} - r_{1n.34\ldots n-1}r_{2n.34\ldots n-1}}{1 - r^2_{1n.34\ldots n-1}} \frac{\sigma_{1.34\ldots n-1}}{\sigma_{2.34\ldots n-1}}.$$

Hence, writing down the similar expression for $b_{21.34\ldots n}$, and taking the square root of the product,

$$r_{12.34\ldots n} = \frac{r_{12.34\ldots n-1} - r_{1n.34\ldots n-1}r_{2n.34\ldots n-1}}{(1 - r^2_{1n.34\ldots n-1})^{\frac{1}{2}}(1 - r^2_{2n.34\ldots n-1})^{\frac{1}{2}}}. \qquad (14)$$

This is, similarly, the expression for $r_{12.n}$ in terms of r_{12}, r_{1n}, and r_{2n}, with the secondary subscripts $_{34\ldots n-1}$ added throughout, and accordingly $r_{12.34\ldots n}$ may be regarded as the partial correlation between $x_{1.34\ldots n-1}$ and $x_{2.34\ldots n-1}$, $x_{n.34\ldots n-1}$ being given, and so on, as for the regression.

15. It is clear that equations (13) and (14) imply a series of relations between correlations or regressions of orders less than $n-2$ with n variables, for all the expressions obtained by eliminating $_{34\ldots n}$ in turn from the secondary subscripts of the constant on the left must be equal to each other. Further, every coefficient of the pth order can be expressed in terms of the coefficients of the $(p-1)$th order in p different ways, by eliminating each of the p secondary subscripts in turn. This enables an absolute check to be kept on the arithmetic by calculating each coefficient in at least two distinct ways.

16. By the use of equation (14), the work of calculating correlation coefficients of higher orders is rendered quite simple and straightforward. The use of equation (13) for calculating the regressions is comparatively

clumsy, however: when the correlations have been found, it is best to work out the standard deviations by equation (11), and then the regressions are given at once by (8). The following data, taken from a discussion of pauperism,* will serve as an arithmetical illustration, the variables being the percentage changes during a decade in the poor-law unions of England in: (1) the percentage of the population in receipt of poor-law relief; (2) the ratio of the numbers given relief out-doors to one indoors (in the workhouse); (3) the proportion of aged (over 60) in the population; (4) the population itself. The values of the correlations of order zero are given in Table I, and the logarithms of $(1 - r^2)^{\frac{1}{2}}$, required in the calculations, are entered in the third column. These coefficients are next grouped in sets of three, one set to each possible group of three variables, as in the second column of Table II, and the coefficients of the first order are then calculated from (14). For convenience in calculating the coefficients of the second order, the values of $\log (1 - r^2)^{\frac{1}{2}}$ are again entered in the last column.

Table I.

Correlation coefficient.		$\log \sqrt{1-r^2}$.
12	+0·52	1·93154
13	+0·41	1·96003
14	−0·14	1·99570
23	+0·49	1·94038
24	+0·23	1·98820
34	+0·25	1·98598

Table II.

Correlation coefficient (zero order).		Product term of numerator.	Numerator.	Correlation coefficient (first order).		$\log \sqrt{1-r^2}$.
12	+0·52	+0·2009	+0·3191	12·3	+0·4014	1·92370
13	+0·41	+0·2548	+0·1552	13·2	+0·2084	1·98071
23	+0·49	+0·2132	+0·2768	23·1	+0·3553	1·94139
12	+0·52	−0·0322	+0·5522	12·4	+0·5731	1·82709
14	−0·14	+0·1196	−0·2596	14·2	−0·3123	1·95544
24	+0·23	−0·0728	+0·3028	24·1	+0·3580	1·94044
13	+0·41	−0·0350	+0·4450	13·4	+0·4642	1·89460
14	−0·14	+0·1025	−0·2425	14·3	−0·2746	1·96595
34	+0·25	−0·0574	+0·3074	34·1	+0·3404	1·94652
23	+0·49	+0·0575	+0·4325	23·4	+0·4590	1·89726
24	+0·23	+0·1225	+0·1075	24·3	+0·1274	1·99290
34	+0·25	+0·1127	+0·1373	34·2	+0·1618	1·98848

* 'Roy. Stat. Soc. Journ.,' vol. 62 (1899), p. 249.

Table III.

Correlation coefficient (first order).		Product term of numerator.	Numerator.	Correlation coefficient (second order).		log $\sqrt{1-r^2}$.
12·4	+0·5731	+0·2131	+0·3600	12·34	+0·458	1·89774
13·4	+0·4642	+0·2630	+0·2012	13·24	+0·276	1·96559
23·4	+0·4590	+0·2660	+0·1930	23·14	+0·266	1·96814
12·3	+0·4014	−0·0350	+0·4364	12·34	+0·458	—
14·3	−0·2746	+0·0511	−0·3257	14·23	−0·359	1·94007
24·3	+0·1274	−9·1102	+0·2376	24·13	+0·270	1·96713
13·2	+0·2084	−0·0505	+0·2589	13·24	+0·276	—
14·2	−0·3123	+0·0337	−0·3460	14·23	−0·359	—
34·2	+0·1618	−0·0651	+0·2269	34·12	+0·244	1·97333
23·1	+0·3553	+0·1219	+0·2334	23·14	+0·266	—
24·1	+0·3580	+0·1209	+0·2371	24·13	+0·270	—
34·1	+0·3404	+0·1272	+0·2132	34·12	+0·244	—

The first order coefficients, from Table II, are then regrouped according to the same primary subscripts as in Table I, and the work repeated precisely as before, as in Table III, but each coefficient of the second order is automatically calculated by this process in two ways and the work thus checked. Small errors introduced by the non-retention of insignificant figures may, of course, prevent complete agreement to the last place of decimals, and for this reason the coefficients of the first order were evaluated to four figures, although only three were required for the final result. In order to obtain the regression equation between changes in pauperism and changes in the three remaining variables, we require the three regressions $b_{12\cdot34}$, $b_{13\cdot24}$, and $b_{14\cdot23}$ and, accordingly, must obtain the six standard deviations, $\sigma_{1\cdot34}$, $\sigma_{2\cdot34}$, $\sigma_{1\cdot24}$, $\sigma_{3\cdot24}$, $\sigma_{1\cdot23}$, $\sigma_{4\cdot23}$.

These are readily calculated and checked by means of the equations of the form—

$$\sigma_{1\cdot34} = \sigma_1 (1-r^2_{13})^{\frac{1}{2}} (1-r^2_{14\cdot3})^{\frac{1}{2}}$$
$$= \sigma_1 (1-r^2_{14})^{\frac{1}{2}} (1-r^2_{13\cdot4})^{\frac{1}{2}}$$

given $\sigma_1 = 29{\cdot}2$, $\sigma_2 = 41{\cdot}7$, $\sigma_3 = 5{\cdot}5$, $\sigma_4 = 23{\cdot}8$; and the values found are :—

$$\sigma_{1\cdot34} = 25{\cdot}61, \qquad \sigma_{1\cdot24} = 27{\cdot}63, \qquad \sigma_{1\cdot23} = 24{\cdot}39,$$
$$\sigma_{2\cdot34} = 36{\cdot}06, \qquad \sigma_{3\cdot24} = 4{\cdot}73, \qquad \sigma_{4\cdot23} = 22{\cdot}86.$$

Hence, from the equations of the form

$$b_{12\cdot34} = r_{12\cdot34} \frac{\sigma_{1\cdot34}}{\sigma_{2\cdot34}},$$

we have

$$b_{12\cdot34} = +0{\cdot}325, \quad b_{13\cdot24} = +1{\cdot}383, \quad b_{14\cdot23} = -0{\cdot}383.$$

That is, the regression equation between changes in pauperism and changes in the other factors considered is

$$x_1 = 0{\cdot}325x_2 + 1{\cdot}383x_3 - 0{\cdot}383x_4.$$

To complete the work, we may calculate $\sigma_{1{\cdot}234}$, the standard error made in estimating x_1 from x_2, x_3, and x_4 by the above equation. The value is

$$\begin{aligned}\sigma_{1{\cdot}234} &= \sigma_1(1-r^2_{12})^{\frac{1}{2}}(1-r^2_{13{\cdot}2})^{\frac{1}{2}}(1-r^2_{14{\cdot}23})^{\frac{1}{2}}\\ &= \sigma_1(1-r^2_{14})^{\frac{1}{2}}(1-r^2_{13{\cdot}4})^{\frac{1}{2}}(1-r^2_{12{\cdot}34})^{\frac{1}{2}}\\ &= 22{\cdot}8.\end{aligned}$$

17. If, in accordance with the notation used for elementary cases in the paper already referred to, and that in a recent note by Mr. R. H. Hooker and myself,* we write

$$\sigma^2_{1{\cdot}23\ldots n} = \sigma^2_1(1-\mathrm{R}^2_{1(23\ldots n)}), \tag{15}$$

$\mathrm{R}_{1(23\ldots n)}$ may be regarded as a coefficient of correlation between x_1 and the expression

$$e_{1{\cdot}23\ldots n} = b_{12{\cdot}34\ldots n}x_2 + b_{13{\cdot}24\ldots n}x_3 + \ldots + b_{1n{\cdot}23\ldots n-1}x_n. \tag{16}$$

The value of R is accordingly a useful datum, as indicating how closely x_1 can be expressed in terms of a linear function of $x_2x_3\ldots x_n$. It may be readily calculated either direct from the equation

$$1-\mathrm{R}^2_{1(23\ldots n)} = (1-r^2_{12})(1-r^2_{13{\cdot}2})\ldots(1-r^2_{1n{\cdot}23\ldots n-1}), \tag{17}$$

or from the value of $\sigma_{1{\cdot}23\ldots n}$ and σ_1, if previously obtained.

It is obvious from (17) that, since every bracket on the right is not greater than unity,

$$1-\mathrm{R}^2_{1\,(23\ldots n)} \leqq 1-r^2_{12}.$$

Hence $\mathrm{R}_{1(23\ldots n)}$ cannot be numerically less than r_{12}. For the same reason, rewriting (17) in every possible form, $\mathrm{R}_{1(23\ldots n)}$ cannot be numerically less than r_{12}, $r_{13}, \ldots r_{1n}$, *i.e.*, any one of the possible constituent coefficients of order zero. Further, for similar reasons, $\mathrm{R}_{1(23\ldots n)}$ cannot be numerically less than any possible constituent coefficient of any higher order. That is to say, $\mathrm{R}_{1(23\ldots n)}$ is not less than the greatest of all the possible constituent coefficients of all orders, and is usually, though not always, markedly greater. Thus in the illustration of § 16, the value of $\mathrm{R}_{1{\cdot}(234)}$ is 0·626, and the greatest correlation coefficient is $r_{12{\cdot}34} = 0{\cdot}458$. The sign of R is necessarily positive, for a positive increment in x_1 obviously corresponds on the average to a positive increment in $e_{1{\cdot}23\ldots n}$. More definitely, the standard deviation of $e_{1{\cdot}23\ldots n}$ is $\sigma_1\mathrm{R}_{1(23\ldots n)}$, and the regression of x_1 on $e_{1{\cdot}23\ldots n}$ is therefore + 1.

Seeing that $\sigma^2_{1{\cdot}23\ldots n} = \sigma^2_1(1 - \mathrm{R}^2_{1(23\ldots n)})$, and that $\sigma_{1{\cdot}23\ldots n}$ is a minimum, we may, alternatively, regard the values of the regressions as determined by the

* 'Roy. Stat. Soc. Journal,' vol. 59 (1906), p. 197.

condition that the correlation between x_1 and $e_{1\cdot 23\ldots n}$, viz., $R_{1(23\ldots n)}$, shall be a maximum.

18. It is obvious that equations (13) and (14) imply relations of an inverse kind, expressing coefficients of a lower order in terms of those of a higher order. Using the same method of expansion as in previous cases, we have

$$\begin{aligned} 0 &= \Sigma(x_{1\cdot 23\ldots n}x_{2\cdot 34\ldots n-1}) \\ &= \Sigma(x_1x_{2\cdot 34\ldots n-1}) - b_{12\cdot 34\ldots n}\Sigma(x_2x_{2\cdot 34\ldots n-1}) \\ &\qquad - b_{1n\cdot 23\ldots n-1}\Sigma(x_nx_{2\cdot 34\ldots n-1}). \end{aligned}$$

That is

$$b_{12\cdot 34\ldots n-1} = b_{12\cdot 34\ldots n} + b_{1n\cdot 23\ldots n-1}b_{n2\cdot 34\ldots n-1}.$$

But by interchanging the suffixes, viz., 1 for n and n for 1,

$$b_{n2\cdot 34\ldots n-1} = b_{n2\cdot 13\ldots n-1} + b_{n1\cdot 23\ldots n-1}b_{12\cdot 34\ldots n-1}.$$

Substituting this value of $b_{n2\cdot 34\ldots n-1}$ in the first equation and simplifying,

$$b_{12\cdot 34\ldots n-1} = \frac{b_{12\cdot 34\ldots n} + b_{1n\cdot 23\ldots n-1}b_{n2\cdot 13\ldots n-1}}{1 - b_{1n\cdot 23\ldots n-1}b_{n1\cdot 23\ldots n-1}}. \tag{18}$$

This is the required equation for the regressions. The similar equation for the correlations is obtained at once by writing down the corresponding expression for $b_{21\cdot 34\ldots n-1}$ and taking the square root

$$r_{12\cdot 34\ldots n-1} = \frac{r_{12\cdot 34\ldots n} + r_{1n\cdot 23\ldots n-1}r_{2n\cdot 13\ldots n-1}}{(1 - r^2_{1n\cdot 23\ldots n-1})^{\frac{1}{2}}(1 - r^2_{2n\cdot 13\ldots n-1})^{\frac{1}{2}}}. \tag{19}$$

19. The general principle that any equation subsisting between such statistical constants as correlations, regressions, and standard deviations holds good for all secondary subscripts, applies also to the equation (3), which expresses the individual deviation of order p in terms of deviations of order zero. That is to say, we have, quite generally, k being any subscript or collection of subscripts,

$$x_{1\cdot 2\ldots kn} = x_{1\cdot k} - b_{12\cdot 3\ldots kn}x_{2\cdot k} - \ldots - b_{1n\cdot 2\ldots k(n-1)}x_{n\cdot k}. \tag{20}$$

For, if l be any one of the subscripts included under k, and if m denote the remaining subscripts, on expanding both sides of the equation in terms of deviations of order zero, the coefficients of $x_1, x_2, \ldots x_n$ are the same. The coefficients of x_l are equal if

$$b_{1l\cdot 2\ldots mn} = b_{1l\cdot m} - b_{2l\cdot m}\,.\,b_{12\cdot 3\ldots kn} - \ldots - b_{nl\cdot m}\,.\,b_{1n\cdot 2\ldots k(n-1)}.$$

But, replacing the regressions $b_{1l\cdot m}$, $b_{2l\cdot m}$, $\ldots b_{nl\cdot m}$ by product sums, this reduces to

$$\Sigma(x_{l\cdot m}\,.\,x_{1\cdot 2\ldots lmn}) = 0,$$

which is true by § 8, whether m denote a single subscript or an aggregate, or is absent, and equation (20) is accordingly correct. Remembering that

$$b_{12\cdot3\cdots kn} = r_{12\cdot3\cdots kn}\frac{\sigma_{1\cdot3\cdots kn}}{\sigma_{2\cdot3\cdots kn}} = r_{12\cdot3\cdots kn}\frac{\sigma_{1\cdot23\cdots kn}}{\sigma_{2\cdot13\cdots kn}},$$

the equation may also be written in the useful form—

$$\frac{x_{1\cdot2\cdots kn}}{\sigma_{1\cdot2\cdots kn}} = \frac{x_{1\cdot k}}{\sigma_{1\cdot2\cdots kn}} - r_{12\cdot3\cdots kn}\frac{x_{2\cdot k}}{\sigma_{2\cdot13\cdots kn}} - \ldots - r_{1n\cdot2\cdots k(n-1)}\frac{x_{n\cdot k}}{\sigma_{n\cdot12\cdots k(n-1)}}. \qquad (21)$$

20. In all the preceding sections no assumption of any kind has been made with respect to the form of the distribution of frequency, but the results may, of course, be applied to the special case of the normal distribution. Let $y_{12\ldots n}$ denote the value of the normal function for the combination of deviations $x_1, x_2, \ldots x_n$, and $y'_{12\ldots n}$ the value of the function when all deviations are zero, then we may write

$$y_{12\ldots n} = y'_{12\ldots n}\cdot\exp -\tfrac{1}{2}\phi\,(x_1x_2\ldots x_n), \qquad (22)$$

the form of the function ϕ being determined by the fact that the distribution of every array must be normal, and that the mean of the array of any one variable associated with given types of the others must be the linear function of those types given by the general regression equation of the form (1). We must have, accordingly

$$\phi = \frac{x^2_1}{\sigma^2_{1\cdot23\ldots n}} + \frac{x^2_2}{\sigma^2_{2\cdot13\ldots n}} + \ldots + \frac{x^2_n}{\sigma^2_{n\cdot12\ldots(n-1)}}$$

$$- 2r_{12\cdot3\ldots n}\frac{x_1x_2}{\sigma_{1\cdot23\ldots n}\sigma_{2\cdot13\ldots n}} - \ldots - 2r_{(n-1)\,n\cdot12\ldots(n-2)}\frac{x_{n-1}x_n}{\sigma_{(n-1)\cdot1\ldots(n-2)}\sigma_{n\cdot1\ldots(n-1)}}. \qquad (23)$$

But this expression may be thrown into several different forms. Thus, replacing the correlated variables, $x_1, x_2, \ldots x_n$, by the independent variables, $x_1, x_{2\cdot1}, x_{3\cdot12}, \ldots x_{n\cdot1\ldots(n-1)}$, we have the very useful form

$$\phi = \frac{x^2_1}{\sigma^2_1} + \frac{x^2_{2\cdot1}}{\sigma^2_{2\cdot1}} + \frac{x^2_{3\cdot12}}{\sigma^2_{3\cdot12}} + \ldots + \frac{x^2_{n\cdot1\ldots(n-1)}}{\sigma^2_{n\cdot1\ldots(n-1)}}. \qquad (24)$$

This expression may be shown to be identical with (23) by expanding in terms of deviations of order zero, and reducing the coefficients of the square terms by means of the equation

$$\frac{r^2_{n\cdot k}}{\sigma^2_{n\cdot k}} + \frac{1}{\sigma^2_n} = \frac{1}{\sigma^2_{n\cdot k}},$$

and those of the product terms by an equation derived at once from (19),

$$\frac{r_{13\cdot2k}r_{23\cdot1k}}{\sigma_{1\cdot3k}\sigma_{2\cdot3k}} - \frac{r_{12}}{\sigma_{1\cdot k}\sigma_{2\cdot k}} = -\frac{r_{12\cdot3k}}{\sigma_{1\cdot3k}\sigma_{2\cdot3k}}.$$

21. Several important results follow at once from the form of the expression (24) for the exponent ϕ. Since the variables are independent, the central value of the normal function, $y'_{12\ldots n}$, must be given by the product of the well-known expressions for the single variables, *i.e.*, we must have

$$y'_{12\ldots n} = \mathrm{N}/(2\pi)^{n/2}\sigma_1\sigma_{2\cdot1}\sigma_{3\cdot12}\ldots\sigma_{n\cdot1\ldots(n-1)}. \qquad (25)$$

22. Again, if we integrate the normal function in the form (24) with respect to x_1, treating the remaining variables, $x_{2\cdot1}$, $x_{3\cdot21}$, etc., as constants of integration, σ_1 is eliminated from y' and x_1 from ϕ, and all the remaining variables in the exponent contain the secondary suffix 1. If $x_{2\cdot1}$, $x_{3\cdot21}$, $\ldots x_{n\cdot1\ldots(n-1)}$ are then replaced by $x_{2\cdot1}$, $x_{3\cdot1}$, $\ldots x_{n\cdot1}$, ϕ may be written in the form (23) for these variables. Similarly, if we integrate again with respect to $x_{2\cdot1}$, $\sigma_{2\cdot1}$ is removed from y' and $x_{2\cdot1}$ from ϕ, and all the remaining variables in the exponent contain both secondary suffixes 1 and 2. If $x_{3\cdot21}$, $x_{4\cdot321}$, $\ldots x_{n\cdot321}$ are then replaced by $x_{3\cdot21}$, $x_{4\cdot21}$, $\ldots x_{n\cdot21}$, ϕ may be written in the form (23) for these variables. Clearly the process may be continued on the same lines. The correlation between all sets of deviations, of any one order, with the same secondary suffixes, is therefore normal correlation.

23. It follows that we may generalise at once the known formulæ for the probable errors of the constants of a normal distribution. Omitting the factor 0·674489...we have, standard error of a

Standard deviation σ_1	$\sigma_1/\sqrt{2\mathrm{N}}$.
Correlation coefficient r_{12}	$1-r^2_{12}/\sqrt{\mathrm{N}}$.
Regression coefficient b_{12}	$\sigma_{1\cdot2}/\sigma_2\sqrt{\mathrm{N}}$.

The first is a well-known result; the last two are cited from the valuable memoir by Professor Karl Pearson and Mr. L. N. G. Filon.* But since $\sigma_{1\cdot k}$ is the standard deviation of the normally distributed variable $x_{1\cdot k}$, $r_{12\cdot k}$ the correlation between the normally distributed variables $x_{1\cdot k}$ and $x_{2\cdot k}$, and $b_{12\cdot k}$ the regression of $x_{1\cdot k}$ on $x_{2\cdot k}$, we must have, quite generally, k denoting as before either a single subscript or an aggregate, standard error of a

Standard deviation $\sigma_{1\cdot k}$	$\sigma_{1\cdot k}/\sqrt{2\mathrm{N}}$.
Correlation coefficient $r_{12\cdot k}$	$1-r^2_{12\cdot k}/\sqrt{\mathrm{N}}$.
Regression coefficient $b_{12\cdot k}$	$\sigma_{12\cdot k}/\sigma_{2\cdot k}\sqrt{\mathrm{N}}$. (26)

The last result may be readily verified against the formula arrived at by Professor Pearson and Mr. Filon, for the case of three variables, after pages of the most laborious work.† The first may be checked for the case of two

* 'Phil. Trans.,' A (1898), vol. 191, p. 229.

† *Loc. cit.*, equation xxxviii, p. 260.

variables, remembering the result of the same writers,* that the correlation between errors in σ_1 and in r_{12} is $r_{12}/\sqrt{2}$; for we have

$$\sigma_{1.2} = \sigma_1(1-r^2_{12})^{\frac{1}{2}};$$

$$\frac{d\sigma_{1.2}}{\sigma_{1.2}} = \frac{d\sigma_1}{\sigma_1} - \frac{r_{12}.dr_{12}}{1-r^2_{12}}.$$

Or, squaring both sides of the equation and summing, using $\epsilon_{1.2}$ to denote the standard error of $\sigma_{1.2}$,

$$\frac{\epsilon^2_{1.2}}{\sigma^2_{1.2}} = \frac{1}{2N} + \frac{r^2_{12}}{N} - \frac{2r_{12}}{1-r^2_{12}}\frac{r_{12}}{\sqrt{2}}\frac{1}{\sqrt{2N}}\frac{1-r^2_{12}}{\sqrt{N}} = \frac{1}{2N}.$$

23. The question of errors of sampling in the case of the coefficient of n-fold correlation, R, is not so simple, owing to the fact that the sign of the coefficient is essentially positive and, consequently, it is subject to biassed error. If, for instance, a series of variables are strictly independent, but values are found for r_{12}, $r_{13.2}$, $r_{14.32}$, etc., equal to δ_2, δ_3, δ_4,...then

$$1-R^2_{1(23\ldots n)} = (1-\delta^2_2)(1-\delta^2_3)\ldots(1-\delta^2_n).$$

If the δ's are sufficiently small to enable us to neglect terms of the fourth order as compared with those of the second order, then we may write to the first approximation,

$$R^2_{1(23\ldots n)} = \delta^2_2+\delta^2_3+\ldots+\delta^2_n.$$

Or, summing for a number of samplings and substituting $1/N$ for $\Sigma(\delta^2)$ in each case, the root-mean-square value of R when the variables are strictly independent is

$$R_0 = (n-1)^{\frac{1}{2}}/N^{\frac{1}{2}}, \qquad (27)$$

n being the number of variables and N the number of observations. R cannot be held with certainty to be of definite significance if not markedly greater than this, and if the number of observations be small compared with the number of variables, the critical value is rather unpleasantly large. Thus in the case of a recent investigation by Mr. R. H. Hooker into the relation between the weather and the crops, $n = 3$, $N = 21$, consequently $R_0 = \sqrt{\frac{2}{21}} = 0{\cdot}31$ (the value cited by him on my authority).† Clearly, if the number of observations be small, it is not worth while dealing with a large number of variables.

* *Loc. cit.*, equation xviii, p. 242.

† 'Roy. Stat. Soc. Journ.,' vol. 70 (1907), p. 7.